CATS Working Group
Internet-Draft
Intended status: Informational
Expires: 13 April 2026

T. Jiang
CMCC
L. Contreras
Telefonica
M. Watts
Verizon
CJ. Bernardos
UC3M
Y. Shang
China Mobile MIGU
P. Liu
China Mobile

10 October 2025

Applicability of CATS Framework
draft-jlmcyp-cats-applicability-00

Abstract

The IETF CATS WG considers the problem of how the network edge can steer traffic between clients of a service and the sites offering the service. The service QoE and/or the performance experienced by edge clients may depend on both network metrics and compute metrics. CATS leverages these metrics and strives to optimize how a network edge node may steer traffic, as appropriate to the service. Revolving around the ’optimized’ objective, the CATS Framework proposes and defines a general architecture for the distribution of network and compute metrics and for the transport of traffic from network edge to service instance. This draft illustrates the applicability of the CATS framework to various noteworthy scenarios.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on 13 April 2026.

Table of Contents

1. Introduction

The CATS WG considers the problem of how the network edge can steer traffic between clients of a service and the sites offering the service. The service QoE and/or the performance experienced by edge clients may depend on both network metrics, such as bandwidth, delay, path loss, reliability, etc., and compute metrics, such as system processing power, storage capacity, system capabilities, etc. CATS leverages these metrics and strives to optimize how a network edge node may steer traffic in a 'balanced' way that is appropriate to the

service.

Revolving around the ’balanced’ objective, the CATS WG has composed three documents, namely, the usecase-requirement draft [CATS-PS-UseCase-Req], the metrics draft [CATS-Metrics-Definition] and the framework draft [CATS.Framework]. Out of the three, the CATS framework draft proposes and defines a general architecture for the distribution of network and compute metrics and for the transport of traffic from network edge to service instance. The CATS framework encompasses various building blocks and emphasizes their interactions, realizing a CATS control and data plane that addresses the ’CATS optimization’ requirements, exploring the distribution scheme of necessary information (e.g., CATS metrics and beyond).

The document revolves around the applicability of the generic CATS framework to some representative scenarios, especially in the mobile and telecom domains.

1.1. CATS Framework

The CATS framework draft [CATS.Framework] standardizes a general CATS architecture that identifies the CATS components along with their interactions, as well as illustrate the workflows of both the control and data planes. The architectural framework facilitates combinationally the making of compute- and network- aware traffic steering decisions in dynamic networking environments with variable computing service resources. Designed as an overlay framework, it guides the selection of the ’suitable’ service (contact) instance(s) from a list of candicates. Note that in the context of CATS, the suitability is subject to the optimal integration of networking and computing metrics.

The CATS framework is comprised of three planes, namely, the management plane, the control plane (CP) and the data plane (DP). Both the CP and DP are more critical, relatively. Further, each plane may consist of respectively several functional elements/ components. E.g., while the CP may be comprised of C-PS, C-SMA, C-NMA, etc., the DP encompasses CATS-forwarder, C-TC, etc.

The clause 3.4 of the CATS Framework [CATS.Framework] illustrates the main CATS functional elements and their interactions, with some of the critical ones described below.

* CATS Service Metric Agent (C-SMA): A control-plane functional component that gathers information about *service* sites and server resources, as well as the status of different service instances. A C-SMA may be standalone-deployed or co-located with other CATS elements.

* CATS Network Metric Agent (C-NMA): A control-plane functional component that gathers information about the state of the underlay *network*. A C-NMA may be implemented as a standalone component or hosted by other CATS component(s).

* CATS Path Selector (C-PS): A control-plane functional componet that utilizes the CATS-domain status (i.e., metrics) collected by C-SMAs and/or C-NMAs to select the egress CATS-forwarder to which the traffic of a given service request is forwarded. A C-PS determines the 'best' path according to both network- and compute-metrics.

* CATS-Forwarder: A data-plane functioinal component (i.e., a network entity) that steers the traffic of a specific service request toward a 'suitable' service (contact) instance based on the combinational effect of network- and compute- metrics. There are two types of them, i.e., the Ingress and the Egress CATS-forwarders.

The above-mentioned critical CATS functional elements and their interacitons are briefly shown in the Figure 1.

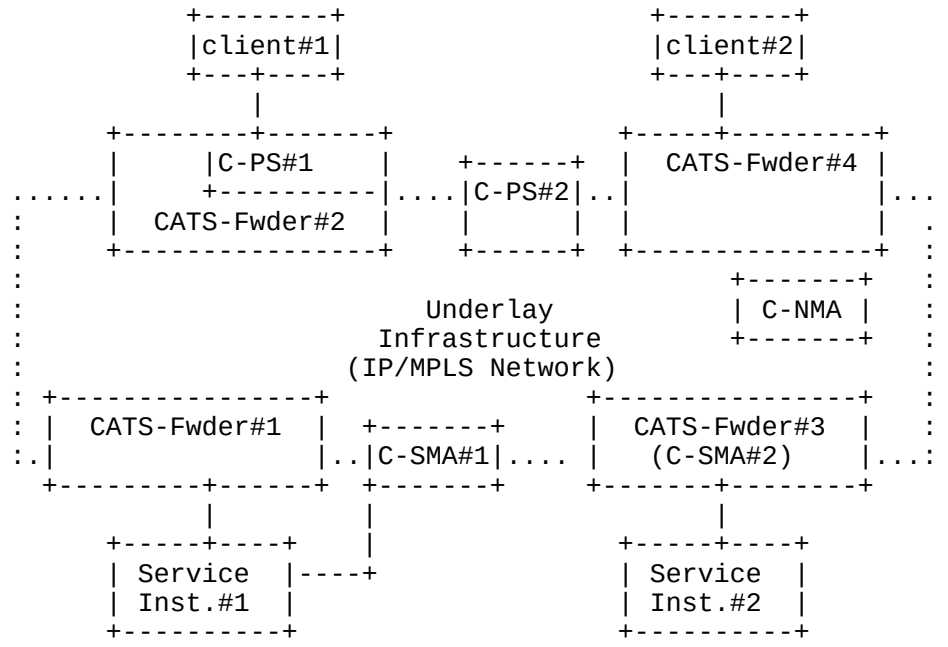

Figure 1: Main CATS Functional Elements (Sketchy)

Note that the 'underlay infrastrucutre' indicates an IP and/or MPLS network that is not necessarily CATS-aware. The CATS paths as computed by a C-PS will be distributed among the CATS-forwarders, which does not impact the underlay nodes. Please reference [CATS.Framework] for more details.

1.2. Applicability of CATS Framework: General

The CATS Framework introduces three deployment options to accommodate a variety of contexts, namely distributed, centralized and hybrid models. However, it does not make any assumption about how the various CATS functional elements are implemented and which deployment model (out of the three) might be adopted. That is, whether a CATS deployment follows the distributed, centralized or even hybrid design is deployment-specific and may only reflect the preferences and policies of the (CATS) service provider. Accordingly, this bodes well for drafting a document to illustrate the applicability of the CATS framework to various noteworthy scenarios, e.g., AI-agent Communication Network or ACN, 5G Edge, and the O-RAN Midhaul, etc.

2. Scenario #1: Applicability of CATS Framework to AI-agent Communication Network (ACN)

2.1. AI-agent Communication Network (ACN)

The CATS ACN draft [CATS.ACN.RefModel] describes the AI-agents along with the network to provide the communication services among various types of AI-agents, i.e., AI-agent Communication Network or ACN.

AI agents are software-driven entities with embedded AI, including ML and NLP, to interact multi-modally with applications, end devices and network components. AI agents play a cruical role in the Telecom domain by enhancing the network efficiency, predicting network conditions, making autonomous & intelligent decisions, and facilitating seamless communication among serviced & servicing entities [CATS.ACN.RefModel]. With the imminently unfolding 6G era, the future world is expected to be full of AI agents, which makes it highly imperative to define a new network framework that is tailored to advance the communication among AI-agents.

This new network framework is defined as the AI-agent Communication Network or ACN. ACN targets at architecting a globally interconnected network to satisfy the on-demand communication, interactions & collaborations with secure and controllable information flow paths for AI-agents in distributed deployment mode. AI-agents demand commonly high computing power and significant energy

consumption, which deems the versatility of the specific realization forms of AI agents. For example, an AI-agent can be instantiated as a standalone physical body, or as an intelligent service (in software state) deployed inside the hosting network, or as a cloud-native instance residing in the edge or remote cloud data centers (DCs), etc.

2.2. Applicability of CATS Framework to ACN

AI-agents in an ACN demand normally intensive compute power and accordingly heavy energy consumptions. Thus, the demands, the capabilities and the processing tasks among AI-agents vary dramatically. It makes more desirable to adopt the seamless collaborations among end devices, networks and (edge or remote) clouds to build a composite AI-agent communication network, for which AI logics are distributed either at the network edge or within the network, either inside or across domains. In that, the compute capabilities at end devices may realize hierarchical AI reasoning with the help of more powerful network entities, which expands the end-side AI services on demand. The network can also provide more advanced AI services to supplement the requirements of (end) AI-agents and lower the compute demands in them. The ultimate goal would be a better balance between achieving the intelligence at AI-agents and the lower energy consumption (of course, potentially more advantages). This certainly conforms to what the CATS is promoting.

The Figure 2 demonstrates the applicability of the CATS framework to the ACN. The figure accommodates the existing C-PS, C-NMA, C-SMA as well as the (new) AI-agents in ACN. The AI-agent#0 is a physical-form AI-agent (e.g., embedded in a server) and the AI-agent#1 is a virtual instance deployed in the cloud service site#1. The service instances #2, #3a and #3b are normal CATS instances deployed in the site#2 and site#3, respectively. The CATS entity C-SMA-2 is in the service site#2 and the C-SMA-3 in the site#3, handling the capture, processing and distribution of compute metrics at service sites. The provider network in the figure contains two CATS network metric agents, i.e., C-NMA-2 and C-NMA-3, for the handling of network metrics. Both C-SMAs and C-NMAs talk to the CATS entity C-PS for metrics distribution.

C-AMA: CATS AI-agent Metric Agnet (May integrate with C-SMA)

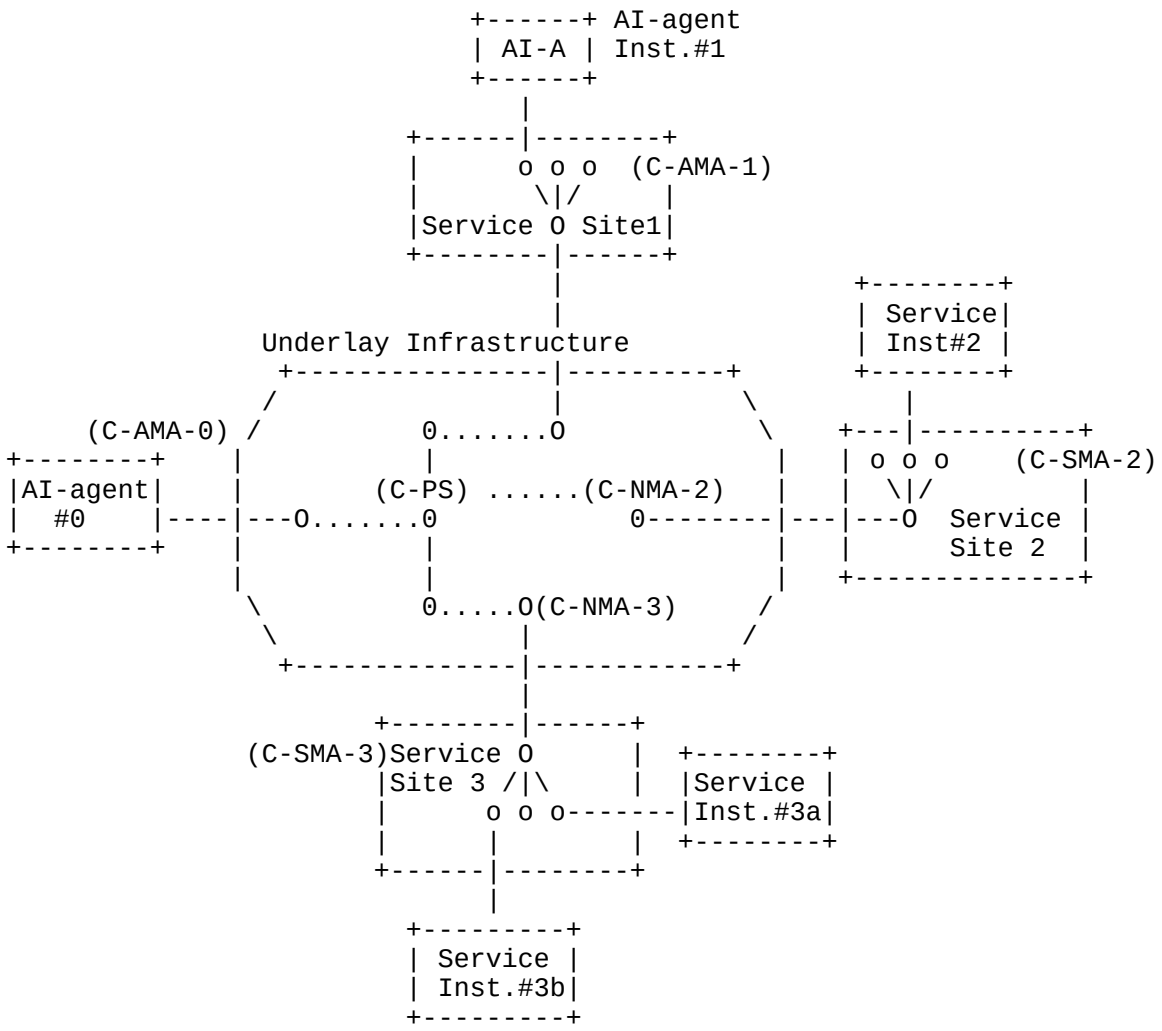

Figure 2: Applicability of CATS Framework to ACN

Note that there is a new type of CATS functional element introduced in the figure, i.e., CATS AI-agent Metric Agent or C-AMA. The C-AMA behaves like the C-SMA, or even being able to be integrated with C-SMA. The C-AMA handles the AI-agent metrics that have been defined in [CATS.ACN.RefModel]. The introduction of C-AMA has no impact on the applicability of the CATS framework to ACN.

3. Scenario #2: Applicability of CATS Framework to 5G Edge Enhancement (5G eEdge)

3.1. 5G Enhanced Edge Computing (5G eEdge)

The 3GPP 5G Edge Computing (EC) enables services to be hosted close to an end device's access point of attachment [TS.23.501] [TS.23.548]. The EC service achieves the efficient delivery through the reduced end-to-end latency and load on the transport network. Edge application servers, or EAS'es, are deployed in (edge) domain networks (DNs) that are connected via the N6 interface of (either central- or local-) UPFs. The 5GC can select either the C-PSA UPF or the L-PSA UPF to optimally foward the UE (uplink) traffic to an EAS with the better (or even the best) 'holistic' metrics.

The 5G enhanced Edge (or eEdge) explores to discover 'suitable' EAS'es to handle edge applications that can be served by multiple EAS'es deployed in different sites. The suitability of an EAS is dependant on both the network metrics, such as bandwidth, latecy, etc., and the compute metrics, such as processing power, storage capacity, AS load, etc. [TR.23.700-49]. Evidently, the integration of both network & compute metrics reflects truly the objectives of the CATS.

Note that, although the 5G eEdge has integrated into the UPF the network metrics (i.e., end-to-end N6 delay over the transport network/TN between (local) PSA UPFs and EAS'es) for optimized EAS selection, the study of the 5G eEDGE has concluded to leave the compute metrics (i.e., load of EAS(es) located in (local) DNs) for further exploration (e.g., in 6G).

3.2. Applicability of CATS Framework to 5G eEdge

This subsection shows how to apply the CATS Framework to 5G eEdge.

In the Figure 3, the UPF-1 to UPF-m indicate 'm' local PSA UPFs, all of which can steer a UE's service request to the suitable application service instance. The applicatioin service or AppService is provisioned in multiple service instances or so-named EAS'es that reside in remote DN(s) or local DN(s), denoted as EAS-1 to EAS-n in the figure. The selection of UPF and EAS depends on the N6 delay, and potentially the EAS load in the future.

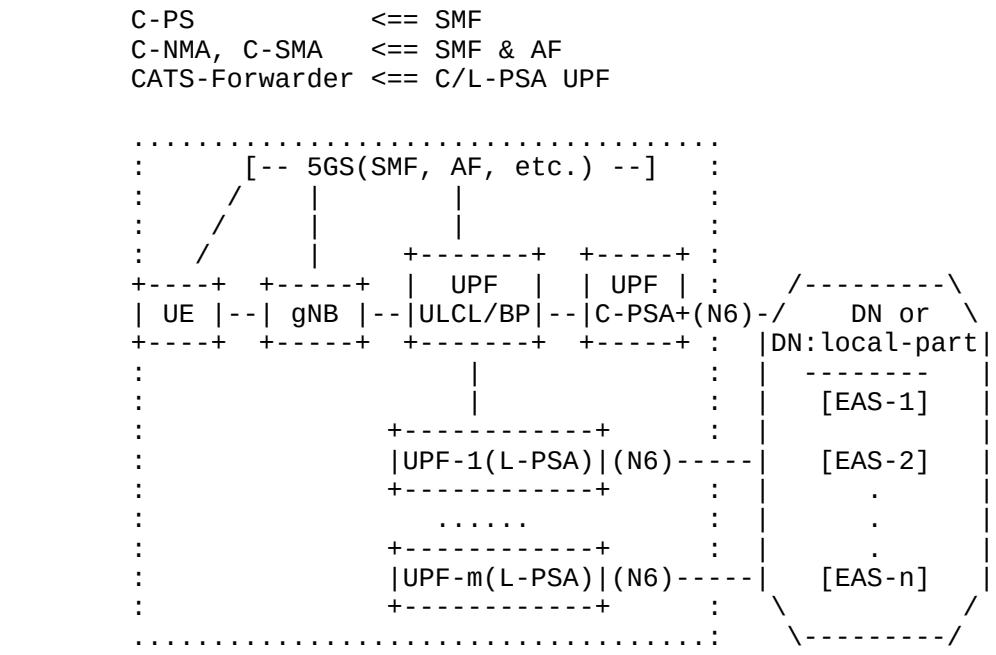

Figure 3: Applicability of CATS Framework to 5G eEdge

Here is the mapping of 5G NFs to CATS functional elements (please reference [CATS.5G.eEdge] for detailed description:

* C-PS vs. SMF: The 5G eEdge has designated a SMF to manage the selection of the optimal UPF (from all candidate UPFs, e.g., UPF-1, ⋯, UPF-m) and the best EAS (from all instances, e.g., EAS-1, ⋯, EAS-n) as in Figure 3) based on the network metric (i.e., the end-to-end N6-delay).

* C-NMA, C-SMA vs. SMF & AF: The combined functionalies of AF & SMF make it a C-NMA. However, because of the non-inclusion of the EAS-load metric, how to map C-SMA is left for future extension (e.g., in 6G).

* CATS-Forwarder vs. C/L-PSA UPF: Upon the policy input from the SMF (i.e., C-PS), UPFs steer the traffic of a service request (via a QoS flow inside a PDU session).

4. Scenario #3: Applicability of CATS Framework to O-RAN Midhaul Networks

The IETF draft [CATS.ORAN.Midhaul] describes the usage of CATS within the Midhaul (MH) networks in the O-RAN architecture. It details how CATS can enhance traffic steering decisions between distributed Units (DUs) and Centralized Units (CUs) by considering both compute metrics (e.g., CPU and memory utilization of CU instances) and network metrics (e.g., bandwidth, latency, reliability of transport networks).

4.1. O-RAN Midhaul

The connection of RU, DU and CU can be performed by means of an IP-based aggregation network. While the FrontHaul (FH) segment connecting RUs and DUs is typically static, the MidHaul (MH) segment connecting DUs and CUs could be more dynamic, subject to the runtime states like system load, workload optimization, energy efficiency, etc. This conforms to the CATS principles to steer the DU-CU flow traffic upon considering both compute and network metrics.

CUs can be deployed in different regions of the network, representing different service instances deployed in distinct service sites. DUs may be running on servers in distinct Data Centers (DCs). Both CUs and DUs are interconnected by an aggregation network (i.e., the IP/MPLS based transport network as assumed in the CATS framework draft [CATS.Framework]).

4.2. Applicability of CATS Framework to Midhaul

As shown in the Figure 4, the PE nodes (being TNEs in O-RAN terminology) play the role of CATS-Forwarders. Each service site is expected to engage with a CATS Service Metric Agent (C-SMA), while the network part is expected to count with a CATS Network Metric Agent (C-NMA). These agents will collect and report metrics to the CATS Path Selector (C-PS), which in this case can be assumed to be part of the TNM in O-RAN terminology (i.e., considering that a centralized deployment model is followed, with the TNM playing the role of centralized control and management element). Example of metrics related to compute could be the CPU average utilization or the memory usage of every CU-UP instance. Please reference the draft [CATS.ORAN.Midhaul] for more details.

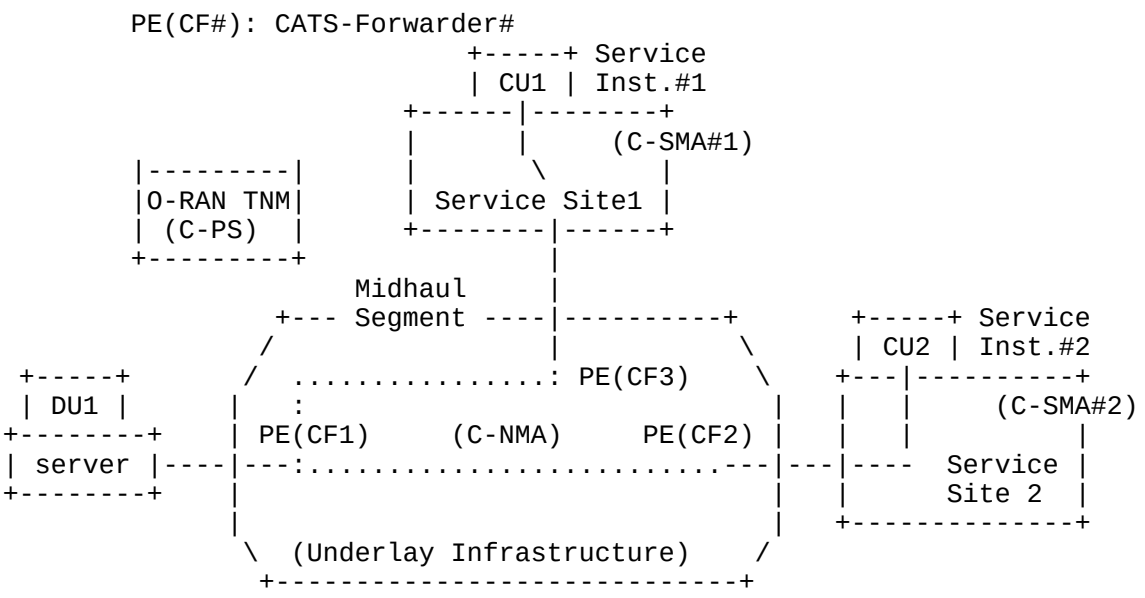

Figure 4: Applicability of CATS Framework to Midhaul

5. Security & Privacy Considerations

The security and privacy considerations follow what have been described in the CATS Framework draft [CATS.Framework].

6. IANA Considerations

There is no IANA requirement.

7. References

7.1. Normative References

[CATS-Metrics-Definition]
Yao, K., et al., "CATS Metrics Definition", draft-ietf-cats-metric-definition, March 2025.

[CATS-PS-UseCase-Req]
Yao, K., et al., "Computing-Aware Traffic Steering (CATS) Problem Statement, Use Cases, and Requirements", draft-ietf-cats-usecases-requirements, June 2025.

[CATS.5G.eEdge]
Jiang, T., et al., "Computing-Aware 5G Edge Enhancement", draft-jiang-cats-usecase-5gedge, October 2024.

[CATS.ACN.RefModel]
Jiang, T., et al., "CATS Reference Model for AI-Agent Communication Network", draft-jiang-cats-reference-acn/, June 2025.

[CATS.Framework]
Li, C., et al., "A Framework for Computing-Aware Traffic Steering (CATS)", draft-ietf-cats-framework, June 2025.

[CATS.ORAN.Midhaul]
Contreras, L., et al., "Compute-Aware Traffic Steering for Midhaul Networks", draft-lcmw-cats-midhaul, July 2025.

[TR.23.700-49]
"TR 23.700-49 v19.0.0: Study on Enhancement of support for Edge Computing in 5G Core network; Phase 3", 3GPP TR 23.700-49, September 2024.

[TS.23.501]
"3GPP TS 23.501 (V19.0.0): System Architecture for 5G System; Stage 2", 3GPP TS 23.501, June 2024.

[TS.23.502]
"3GPP TS 23.502 (V19.0.0): Procedures for the 5G System; Stage 2", 3GPP TS 23.501, June 2024.

[TS.23.548]
"5G System Enhancements for Edge Computing; Stage 2", 3GPP TS 23.548, June 2025.

7.2. Informative References

Authors' Addresses

Tianji Jiang
CMCC
Email: tianjijiang2012@gmail.com

Luis M. Contreras
Telefonica
Email: luismiguel.contrerasmurillo@telefonica.com

Mark Watts
Verizon
Email: mark.t.watts@verizon.com

Carlos J. Bernardos
UC3M
Email: cjbc@it.uc3m.es

Yuxiang Shang
China Mobile MIGU
Email: shangyuxiang@migu.chinamobile.com

Peng Liu
China Mobile
Email: liupengyjy@chinamobile.com